UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Machine-Learning-aided indicator extraction from HPV (Human Papilloma Virus) & cancer-related medical articles

A Degree Thesis

Submitted to the Faculty of the

Escola Tècnica d'Enginyeria de Telecomunicació de Barcelona

Universitat Politècnica de Catalunya

by

Marc Armenter Hierro

In partial fulfilment

of the requirements for the degree in

TELEMATICS ENGINEERING

Advisor: Alfonso Rojas

Barcelona, October 2019

Abstract

The idea of the project is based on the initiative to provide a solution regarding data extraction from text, specially from medical articles, to collaborate in developing a tool that automates or facilitates the detection of certain indicators. In this project we will focus on detecting, in an article, the number of subjects under study, concretely, subjects referring to human papilloma virus and others related to cancer.

During the project, several techniques will be developed and the results will be analyzed to determine which methods are the most accurate to deal with the problem. All in an open source code written in Python that can manage to introduce variants of analysis.

The results have varied throughout the project, and it has been concluded that to extract information correctly from the text, the most important thing is to extract the maximum characteristic information from the positive cases, in such a way that their class is more clearly defined, regardless of the method of classification later used. That is, it is not always necessary to use the most efficient method of the market if we are able to characterize our information in such a way that it is more easily separable.

In our particular case, we have seen that the more we have been narrowing down and deepening in extracting relevant information, better results have emerged. Regarding the nominal phrases where we find positive cases, we have been able to mount a structure that has detected X% of the indicators despite having an accuracy of X%, which means that we are detecting the majority of positive cases, but we are adding negative candidates to the group classified as positive.

It should be mentioned that if you have a large database with a high capacity calculation machine, in most cases where the language appears, the neural networks will eventually be the most sensitive and accurate; given the fact that language contains a lot of intrinsic manipulable and relevant information that can be taken unto account when it comes to characterizing a sample. But to achieve consitency, many cases are necessary with examples where words are used in different contexts in order to create dimensions to represent them.

Resum

La idea del projecte neix de la iniciativa d'aportar una solució pel que fa a l'extracció de dades de text, concretament en articles mèdics, per col·laborar a desenvolupar una eina que automatitzi o faciliti la detecció de certs indicadors. En el projecte ens centrarem en detectar, en un article, el nombre de subjectes sotmesos a proves, més contretament, subjectes referents al virus del papil·loma humà i d'altres relacionats amb el càncer.

Durant el projecte es desenvoluparan varies tècniques i s'analitzaran els resultats per determinar quines tècniques funcionen millor per afrontar el problema. Tot en un codi de descàrrega pública escrit en Python que està pensat per introduir variants d'anàlisis.

Els resultats han anat variant al llarg del projecte, i s'ha pogut concloure que per extreure informació correctament del text, el més important és aconseguir extreure la màxima informació característica dels casos positius, de tal manera que la seva classe quedi més ben definida, independentment del mètode de classificació posteriorment utilitzat. És a dir que no sempre caldrà usar el mètode més eficient del mercat si som capaços de caracteritzar la nostra informació de tal manera que sigui més fàcilment separable.

En el nostre cas particular, hem vist que com més hem anat acotant i aprofundint en l'extracció de informació rellevant, més bons resultats han aparegut. Fixant-nos en els sintagmes nominals on hi trobem els casos positius, s'ha pogut muntar una estructura que ha acabat per detectar un 75% dels indicadors tot i tenir una precisió del 50%, cosa que significa que estem detectant la majoria de casos positius, però estem afegint candidats negatius al grup classificat com positius.

Cal recalcar que si es disposés d'una immensa base de dades amb una màquina d'alta capacitat de càlcul, en la majoria de casos on el llenguatge aparegui, les xarxes neuronals acabaran resultant les més sensibles i precises; donat pel fet que el llenguatge conté un munt de informació intrínseca que és possible manipular i tenir en compte a l'hora de caracteritzar una mostra, però que per ser capaç de donar-li forma, són necessaris molts casos on s'utilitzin les paraules en diferents contextos per tal de crear dimensions on representar-les.

Resumen

La idea del proyecto nace de la iniciativa de aportar una solución con respecto a la extracción de datos de texto, concretamente en artículos médicos, para colaborar a desarrollar una herramienta que automatice o facilite la detección de ciertos indicadores.

Durante el proyecto nos centraremos en detectar, en un artículo, el número de sujetos sometidos a pruebas, más contretamente, sujetos referentes al virus del papiloma humano y otros relacionados con el cáncer.

Durante el proyecto se desarrollarán varias técnicas y analizarán los reultats para determinar qué métodos son los más precisos para afrontar el problema. Todo en un código de descarga pública escrito en Python que está pensado para introducir variantes de analisis.

Los resultados han ido variando a lo largo del proyecto, y se ha podido concluir que para extraer información correctamente del texto, lo más importante es conseguir extraer la máxima información característica de los casos positivos, de tal manera que su clase quede mejor definida, independientemente del método de clasificación posteriormente utilizado. Es decir, que no siempre será necesario usar el método más eficiente del mercado si somos capaces de caracterizar nuestra información de tal manera que sea más fácilmente separable.

En nuestro caso particular, hemos visto que cuanto más hemos ido acotando y profundizando en la extracción de información relevante, mejores resultados han aparecido. Fijándonos en los sintagmas nominales donde encontramos los casos positivos, se ha podido montar una estructura que ha terminado para detectar un 75% de los indicadores a pesar de tener una precisión del 50%, lo que significa que estamos detectando la mayoría de casos positivos, pero estamos añadiendo candidatos negativos al grupo clasificado como positivos.

Hay que recalcar que si se dispusiera de una inmensa base de datos con una máquina de alta capacidad de cálculo, en la mayoría de casos donde el lenguaje aparezca, las redes neuronales acabarán resultando las más sensibles y precisas; dado por el hecho de que el lenguaje contiene un montón de información intrínseca que es posible manipular y tener en cuenta a la hora de caracterizar una muestra, pero para ser capaz de darle forma, son necesarios muchos casos donde se utilicen las palabras en diferentes contextos para crear dimensiones donde representarlas.

Agraïments

Després de l'elaboració d'aquest projecte, només resta agrair el suport de totes les persones i amics que m'han ajudat a fer possible aquest treball de fi de grau.

Especialment, vull donar les gràcies a l'Alfonso per la seva confiança i suport al llarg del projecte; al David per la seva paciència i per ajudar-me a resoldre tots els dubtes, que no han estat pocs; i per suposat als meus pares, a qui dedico aquest treball, pel seu suport incondicional i per donar-me l'oportunitat d'arribar fins aquí.

Historial de revisions i registres de validació

Revision	Date	Purpose
0	18/06/2019	Document creation
1	19/07/2019	Document revision
2	28/08/2019	Document revision
3	20/09/2019	Document revision
4	05/10/2019	Document approval

DOCUMENT DISTRIBUTION LIST

Name	e-mail
Marc Armenter Hierro	marcarmeter@hotmail.com
Alfonso Rojas	david.gomez.guillen@gmail.com
David Gòmez	david.gomez.guillen@gmail.com

Written by:		Reviewed and approved by:	
Date	05/10/2019	Date	07/10/2019
Name	Marc Armenter Hierro	Name	Alfonso Rojas
Position	Project Author	Position	Project Supervisor

Taula de continguts

Llistat de Figures

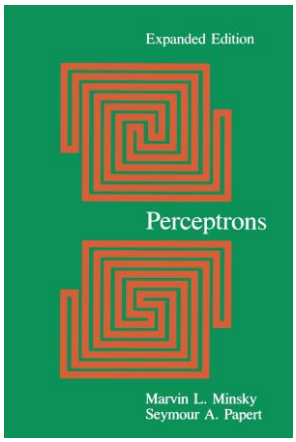


Fig. 1: Perceptron, Marvin Minsky (pàg. 25)

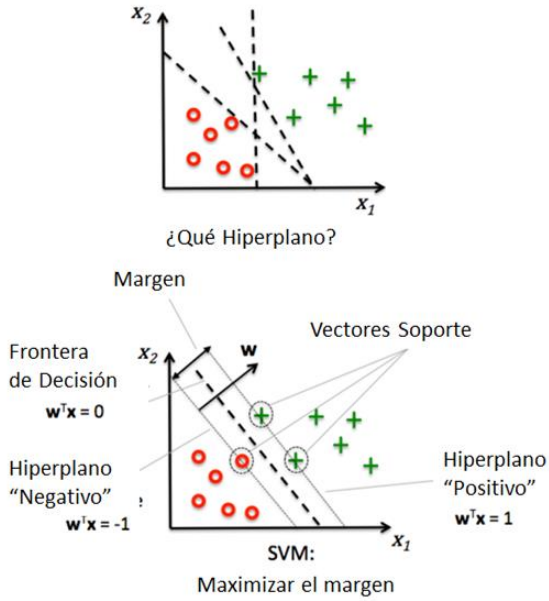


Fig. 2: SVM (pàg. 33)

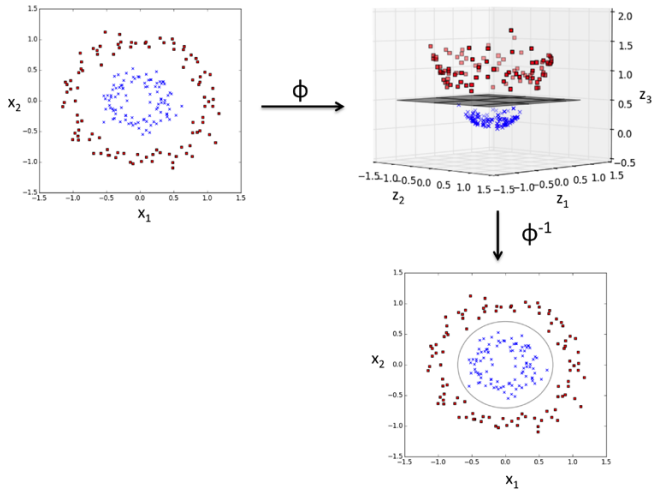


Fig. 3: Kernelització en SVM (pàg. 33)

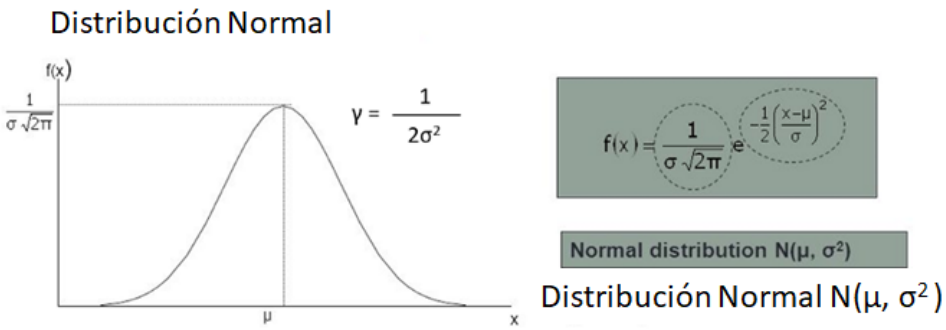


Fig. 4: El paràmetre gamma i la seva relació amb la distribució normal (pàg. 33)

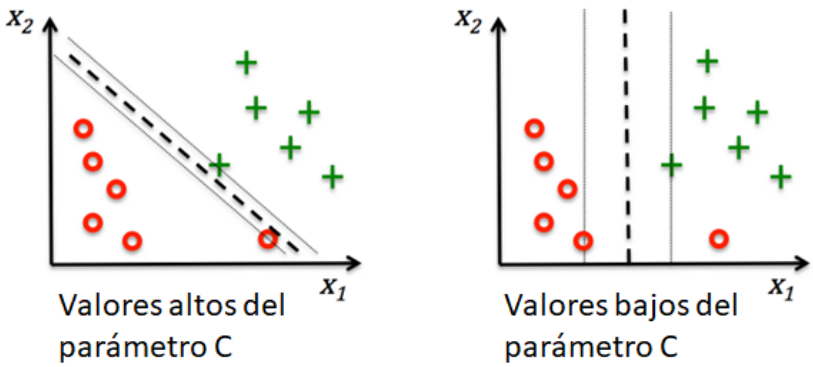


Fig. 5: El paràmetre C i el seu comportament (pàg. 33)

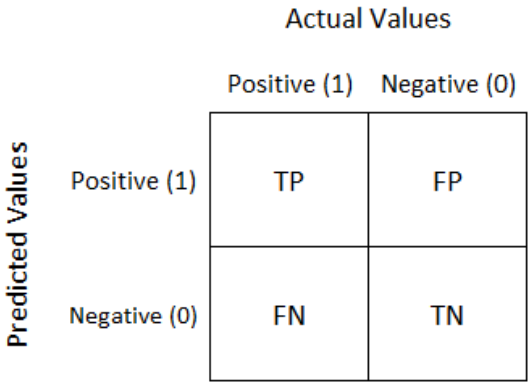


Fig. 6: Matriu de confusió (pàg. 34)

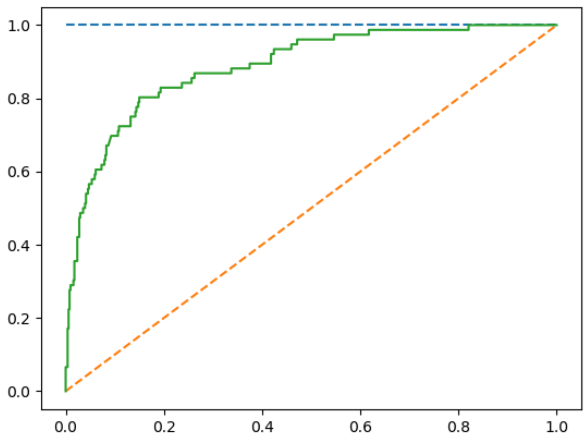


Fig. 7: ROC dades corrompudes (pàg. 36)

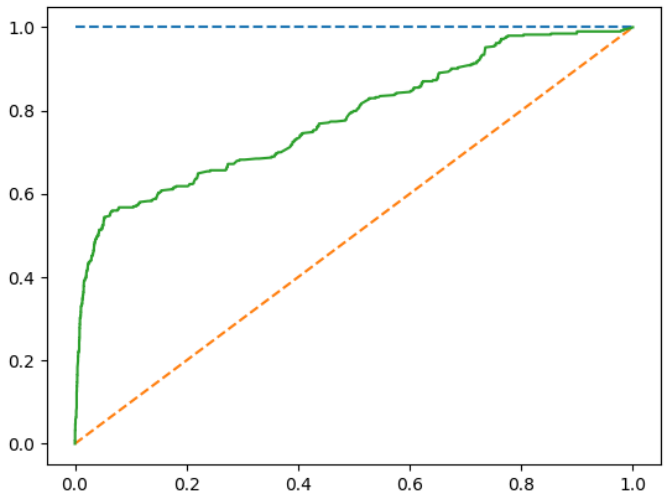


Fig. 8: ROC SVM amb CV (pàg. 36)

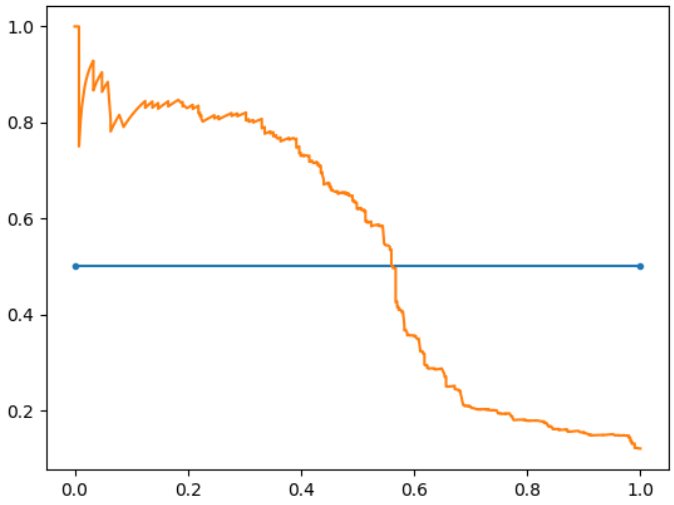


Fig. 9: Sensibilitat-Precisió SVM amb CV (pàg. 36)

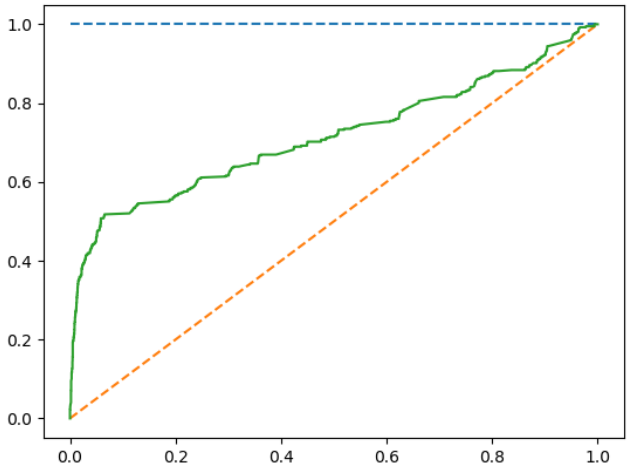


Fig. 10: ROC SVM amb CV i distàncies (pàg. 36)

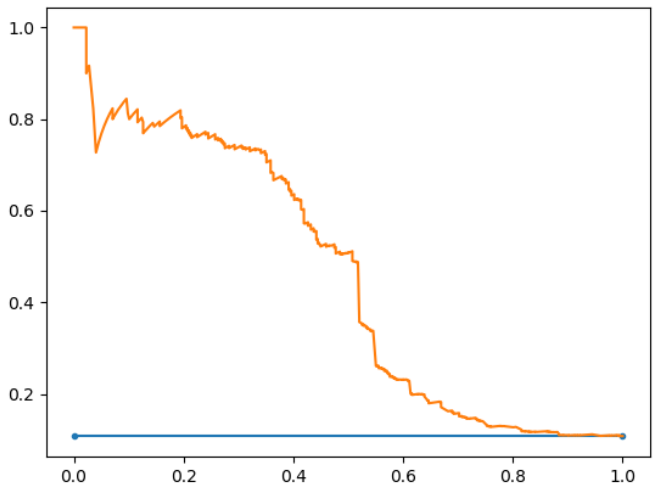


Fig. 11: Sensibilitat-Precisió SVM amb CV i distàncies (pàg. 36)

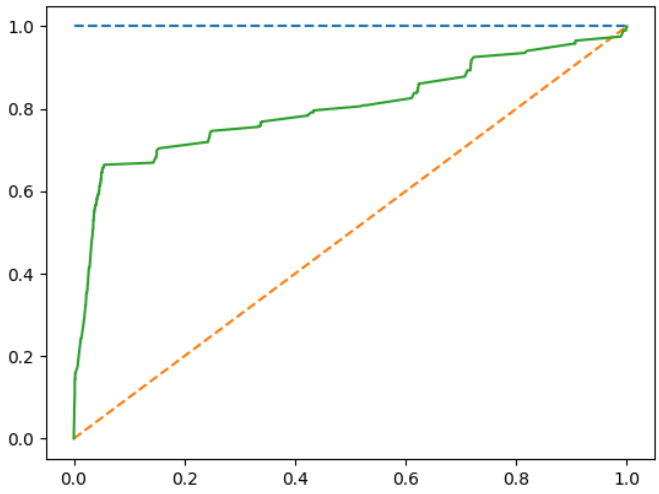


Fig. 12: ROC SVM amb CV i NLP (pàg. 36)

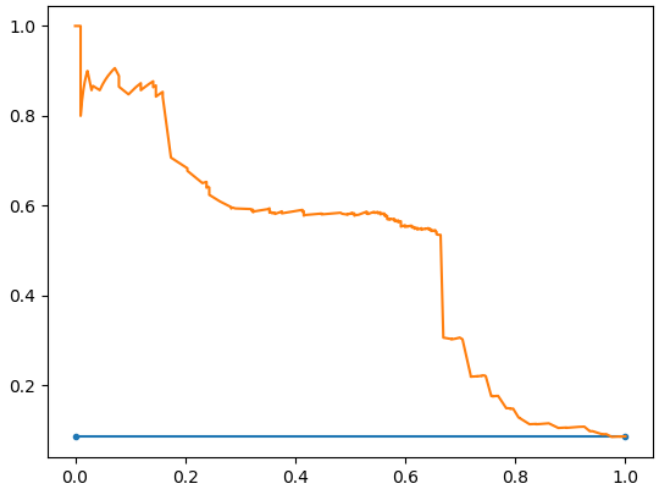


Fig. 13: Sensibilitat-Precisió SVM amb CV i NLP (pàg. 36)

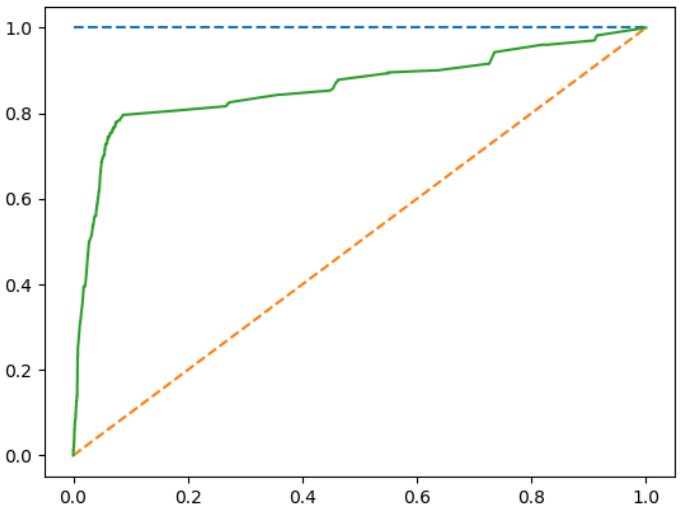


Fig. 14: ROC SVM balancejat amb CV i NLP (pàg. 37)

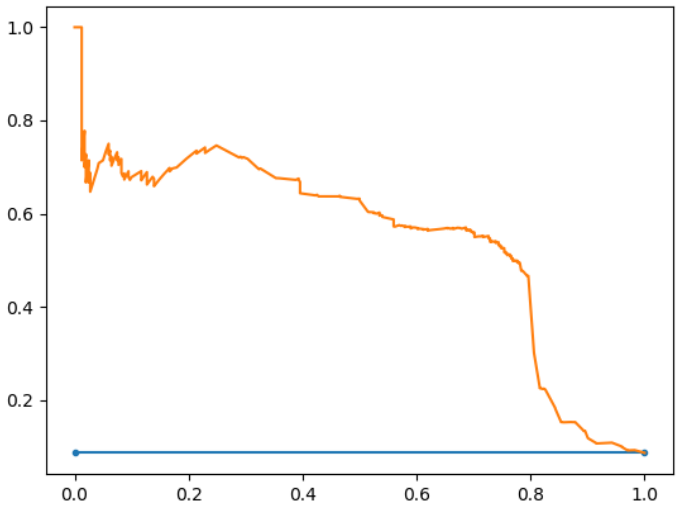


Fig. 15: Sensibilitat-Precisió SVM balancejat amb CV i NLP (pàg. 37)

Llistat de Taules

pmid	author	year	n	n_ic
16810142	Abdel Aziz	2006	166	156
20716343	Adjorlolo-Johnson	2010	132,120	110
19152401	Agarossi	2009	9947	9148
24977385	Agorastos	2014	5107	5107
25712774	Aguilar-Lemarroy	2015	822	356
24192311	Akarolo-Anthony	2013	278	108
23534783	Akcali	2013	410	410
24619263	Al-Ahdal	2014	519	455
10530768	Alexandrova	1999	309	309
21252451	Alhamany	2010	938	785

Taula 1: Porció de dataframe inicial (pàg. 30)

PMID	DateCompleted	DateRevised	JournalTitle	JournalCountry	KeywordList	MeshHeadingList	Lang	PublicationDate	ArticleTitle	Abstract	Authors	Candidate
16810142	2006/10/25	2006/11/15	Medical science monito	United States	[]	Adult, DNA, Viral, Egy	eng	2006 Jul	Screening for	HPV infect	Mohamed Ta	166
20716343	2010/12/13	2018/11/13	BMC infectious diseases	England	[]	Adolescent, Adult, Ag	eng	2010 Aug	Assessing the	The associ	Georgette Ad	132,120
19152401	2009/02/16	2009/01/22	Journal of medical virol	United States	[]	Adolescent, Aged, Ce	eng	2009 Mar	Prevalence ar	The aim of	Alberto Agar	9947
24977385	2015/03/30	2014/07/29	European journal of can	England	[]	Adult, Cross-Sectiona	eng	2014 Sep	Epidemiology	The object	Theodoros A	5107
25712774	2015/12/03	2015/03/18	Journal of medical virol	United States	[ListElement([StringEl	Adolescent, Adult, Ag	eng	2015 May	Human papill	The preval	Adriana Aguil	822
24192311	2014/11/21	2019/03/18	BMC infectious diseases	England	[]	Adult, Coinfection, Fe	eng	2013 Nov	HIV associate	In develop	Sally N Akaro	278
23534783	2013/12/03	2019/06/06	Asian Pacific journal of	Thailand	[]	Adult, Cervix Uteri, Co	eng		Human papill	To determ	Sinem Akcali,	410
24619263	2014/10/24	2018/10/23	Journal of infection in d	Italy	[]	Adult, Aged, Cytologic	eng	2014 Mar	Human papill	Certain ge	Mohammed	519
10530768	1999/11/05	2019/07/20	Cancer letters	Ireland	[]	Adolescent, Adult, Fe	eng	1999 Oct	Features of H	Prevalence	Y N Alexandr	309
21252451	2011/05/12	2013/11/21	Journal of infection in d	Italy	[]	Adolescent, Adult, Ag	eng	2010 Nov	Prevalence of	Many stud	Zaitouna Alh	938
21292583	2012/01/19	2013/11/21	Cancer epidemiology	Netherlands	[]	Adolescent, Adult, Ce	eng	2011 Oct	Human papill	No accura	Tamar Alibeg	1309
17977997	2008/03/18	2018/11/13	Journal of clinical micro	United States	[]	Adult, Female, Humar	eng	2008 Feb	Cervical huma	The preval	Bruce Allan,	1073
17437272	2007/09/12	2016/03/03	International journal of	United States	[]	Adult, Alphapapilloma	eng	2007 Aug	Cervical scree	Cervical ca	Maribel Almo	5435
18708390	2008/12/16	2008/08/18	Cancer epidemiology, bi	United States	[]	Adolescent, Chi-Squa	eng	2008 Aug	Prevalence of	Infection v	Pietro Amma	1006
12878093	2003/11/10	2006/11/15	Journal of clinical virolo	Netherlands	[]	Adult, Aged, Aged, 80	eng	2003 Aug	Molecular det	Uterine ce	Mariam Amr	147,447
12655524	2003/04/16	2006/11/15	Cancer	United States	[]	Adult, Carcinoma, Ce	eng	2003 Apr	Correlation o	Human pa	Hee Jung An,	1983,1650
21603766	2011/11/03	2019/05/17	Revista panamericana d	United States	[]	Adolescent, Adult, Ag	eng	2011 Apr	Human papill	Human pa	Glennis M An	310
15950365	2005/11/07	2013/11/21	European journal of obs	Ireland	[]	Adult, Age Distributio	eng	2005 Jul	Prevalence of	To study th	Raksha Arora	3300

Taula 2: Porció de dataframe amb abstractes (pàg. 30)

Is fraction	N	N close words	N close words distances	N sentence words	PMID	Is Candidate
0	18	['HPV', 'type']	[1, 2]	['The', 'second', 'generation', 'HC', 'II', 'te	16810142	0
0	166	['Egyptian', 'woman']	[1, 2]	['We', 'evaluate', 'Egyptian', 'woman']	16810142	1
1	166	[]	[]	['The', 'overall', 'prevalence', 'of', 'HPV',	16810142	1
0	25	['HPV', 'positive', 'wom	[1, 2, 3]	['Among', 'the', 'HPV', 'positive', 'woman	16810142	0
0	16	[]	[]	['Among', 'the', 'HPV', 'positive', 'woman	16810142	0
0	64	['%']	[1]	['Among', 'the', 'HPV', 'positive', 'woman	16810142	0
0	4	[]	[]	['Among', 'the', 'HPV', 'positive', 'woman	16810142	0

Taula 3: Porció de dataframe amb tots els candidats extrets (pàg. 33)

1. Introducció

1.1. Orígen del projecte

La presa de decisions ha estat sempre un afer lligat als humans, les prenem des de bon matí fins que anem a dormir i sempre estan presents en el nostre dia a dia. És tant així que malgastem un munt d'energia en prendre les petites decisions diàries, decisions ràpides i de vegades banals, però que fan treballar el nostre cervell. De fet, quan prenem decisions, mai podrem tenir en compte totes les variables que afecten una situació per saber quina és l'opció més correcte o encertada, ja que la nostra condició evolutiva no ens ha permès desenvolupar aquesta capacitat de contemplar tantes possibilitats simultàniament. El que tendim a fer és a seguir un model més aviat senzill que ens proporciona el nostre cervell per saber quina acció serà la més adient, un model que acostuma, tot i a servir-nos per sobreviure o prendre decisions ràpides, a ser de tot menys rigorós; probabilísticament parlant.

És aquí on entra en joc la capacitat computacional que tenen per oferir els ordinadors. Juntament amb els algoritmes que han sortit a la llum les últimes dècades, s'ha pogut desenvolupar una tecnologia basada en un aprenentatge automàtic que, amb les dades adients i una bona quantitat d'exemples, és capaç de fer prediccions i encertar amb alt grau de confiança.

Poc a poc s'han anat introduint aquestes tècniques en varis camps com en sistemes de recomanació, sistemes conversacionals, classificació d'imatges i construcció de mapes de profunditat, assistents de veu, intel·ligència artificial de jocs, reconeixement facial o d'impremta i una llista sense fi de possibles aplicacions. Algunes d'aquestes aplicacions han ajudat un munt a optimitzar processos tant importants com el diagnòstic o la detecció de malalties en pacients, casos que han salvat vides. Però tot i la immersió de les noves tecnologies en el sector de la medicina, encara queden molts aspectes on aquestes poden aportar un cop de mà i estalviar, a experts en el camp mèdic, un munt de temps a dedicar en tasques tedioses; com són els casos de la lectura i classificació d'articles o el d'emplenar i validar formularis.

És d'aquí d'on parteix la idea del projecte. L'Institut Català d'Oncologia (ICO) juntament amb l'Agència Internacional de Recerca sobre el Càncer (IARC, de l'anglès International Agency of Research on Cancer) tenen un centre d'informació que tracta el Virus del Papil·loma Humà (HPV, de l'anglès Human Papiloma Virus) a Barcelona, i des de fa un temps, tenen la idea de crear un model d'aprenentatge automàtic que sigui capaç de detectar certa informació en articles mèdics de la seva base de dades. Donada aquesta circumstància, ha estat concedida la iniciativa d'iniciar aquest projecte des de zero i amb l'ajut del supervisor de l'ICO David Gòmez.

Al ser un tema poc tractat, s'ha fet la proposta inicial de desenvolupar aquest model per, concretament, detectar el nombre de persones, mostres o casos que han estat sotmesos a l'estudi d'aquell mateix article.

1.2. Objectives

Com veurem més endavant en la presentació del projecte, l'extracció d'informació de dades en format de text, no té massa recorregut pel que fa a reconeixement d'informació concreta, com és el cas del projecte. És per això que la fita establerta no té una base ferma en la que fer referència, només alguns pocs estudis[1][2].

L'objectiu marcat a l'inici del projecte va ser d'encertar en un 80% dels articles i amb un 70% de probabilitat de que l'indicador candidat sigui de la classe positiva. Cal deixar clar que aquest és un objectiu que supera les expectatives reals del projecte i que s'ha establer per tenir una meta, així que només que ens acostem a aquests valors, ja es podrà considerar un bon resultat.

A més, però, l'objectiu del projecte no és només construir un model, sinó una aplicació prototip que serveixi per generar i entrenar models i fer anàlisis per poder-los guardar per usar-los més endavant, és a dir, marcar un punt de partida per poder seguir incorporant utilitats en un futur sense haver de fer massa esforços de codi.

D'aquesta manera es podran incorporar diferents models i comparar-los entre ells amb diferents mètriques.

1.3. Eines i metodologia

Les eines principals per dur a terme el projecte són eines de software, concretament llibreries de Python d'ús conegut en el sector. No totes les llibreries usades al llarg del projecte es nombraran aquí, només les més importants i específiques per desenvolupar el projecte.

Com que només es disposa d'una base de dades on apareixen els identificadors d'articles amb els nombres de candidats bons (nombre d'afectats/estudiats) per cada article, per començar serà necessària la descàrrega d'aquests per tal de manipular-los. Per fer aquesta tasca disposem d'Entrez, una llibreria que ofereix una API REST que ens permet, a través d'HTTP, accedir a la base de dades del National Center for Biotechnology Information (NCBI), la qual recull una immensa llista de literatura mèdica, seqüencies dels gens, genomes, proteïnes i altra informació d'utilitat en els camps d'estudi biotecnològics.

Un cop els tinguem descarregats, podrem iniciar un preprocessat dels articles, cosa que podrem aconseguir amb la llibreria NLTK, la qual integra funcions pel tractament el text. A l'haver localitzat i tractat les dades amb aquest primer retoc, podrem passar ja a un primer pas d'entrenament amb la llibreria SKLEARN. Aquesta llibreria ens posa a disposició un gran ventall de mètodes d'aprenentatge; bàsicament usarem aquesta llibreria per gairebé tots els models d'aprenentatge automàtic, tant per transformar les paraules en valors/pesos en un primer pas, com per classificar les classes positiva i negativa posteriorment.

Finalment, un cop el classificador obtingui les prediccions, podrem fer ús del mòdul integrat a SKLEARN anomenat METRICS, la qual ens permet treure estadístiques

basant-se amb els resultats de les prediccions del model, juntament amb la llibreria MATPLOTLIB, la qual ens servirà per mostrar gràfics, com les corbes ROC del model en qüestió, més endavant veurem la seva utilitat.

Tot aquest codi, s'anirà penjant a GitHub en un directori de l'ICO, on es podrà descarregar i utilitzar lliurament.

El link és el següent: https://github.com/hpv-information-centre/ml-data-extractor.

A més, totes les referències de les llibreries amb la seva documentació es pot trobar a les corresponents pàgines web amb una extensa descripció del seu funcionament.

1.4. Incidències i desviacions

En un inici es va designar un temps a la recerca, temps que es va quedar curt al veure que no hi havia una ruta marcada per resoldre el problema i s'havia d'enginyar alguna solució al problema. La literatura al respecte és escassa i en poques ocasions ens parla sobre la detecció de qualitats o indicadors en text pla, i quan ho fa, són indicadors molt diferents als del projecte i normalment, més fàcils d'identificar (ja sigui perquè s'identifiquen amb símbols o en textos encara més estructurats que els mèdics). Les tècniques que més s'han explotat fins al dia d'avui focalitzen els seus esforços en detectar una intencionalitat o un missatge emotiu en les paraules usades per transmetre-ho i que la màquina acabi per aprendre el llenguatge. Per aquesta raó, el tema del projecte queda relativament marginat. I tot i que la ruta del procediment a seguir estava marcada i parlada amb el supervisor, feia falta tenir un recolzament literari que justifiqués o respongués per la metodologia del projecte. I és per això que al trobar informació sobre la temàtica del projecte va caldre adaptar alguns passos, prolongant el desenvolupament.

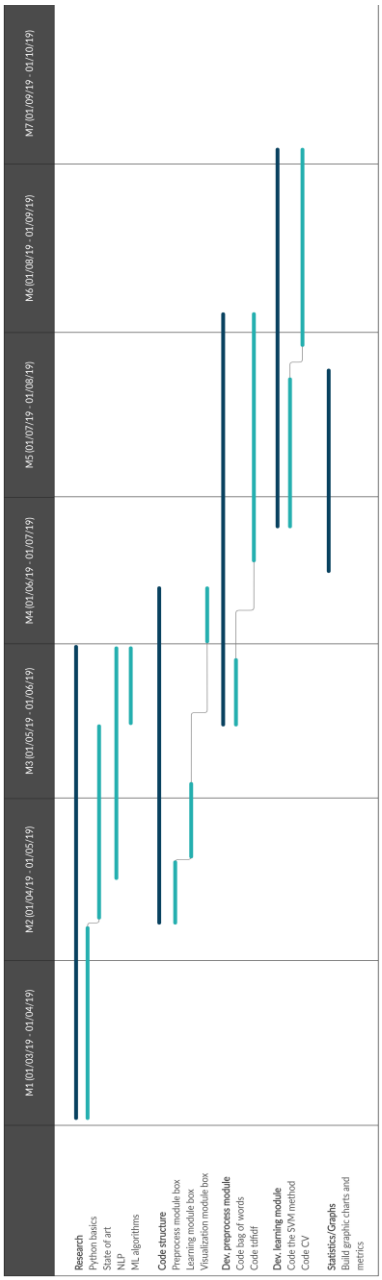
A més, es van fer proves un cop tot el procés estava ben encadenat, i aquestes van tenir uns resultats inesperats. Havien estat sorprenentment bons, cosa que va cridar l'atenció i va fer desconfiar de la seva veridicitat. I efectivament, hi havia un afer que no s'havia tingut en compte: per tal de que el model sigui més rellevant, és a dir, que tingui més alta fiabilitat, no es pot usar mai una mostra de test per entrenar el model, i el procés que estava operant en aquell moment, usava tots els articles per trobar una estructura característica dels indicadors. En altres paraules, estàvem tenint en compte tots els articles que teníem a disposició per extreure'n informació que ens fos útil per classificar en les següents etapes.

En aquest punt es va decidir implementar la cross-validation, la qual ens permet obtenir una estadística mitja i més fiable de les mostres; sobretot en els casos en que no es tenen tantes mostres amb les que treballar. La llibreria de Python que implementa aquests mètodes, incorpora una cross-validation per un mètode de classificació, però el nostre mòdul de preprocessat també necessitava tenir en compte aquest fet. És per això que es va decidir implementar una cross-validation personalitzada i més específica pel tractament que es volia implementar.

Finalment, es planejava tenir una pàgina web muntada per la finalització del projecte on es pogués fer una demostració i es poguessin veure estadístiques generades pel model, però l'agenda ha estat més atapeïda del planejat i és per això que es posposarà i

s'implementarà cara la defensa del projecte. La intenció serà poder veure en directe un test simple i explicar què ens volen dir les mètriques i gràfics en detall.

1.5. Pla de projecte final



Work Packages:

<table>
<tr><td>Project: Research</td><td colspan="2">WP ref: 1</td></tr>
<tr><td>Major constituent: Investigate and learn</td><td colspan="2">Sheet 1 of 5</td></tr>
<tr><td rowspan="2">Short description:
Focus on acquire the necessary knowledge such as what is and how to implement NLP over text or how machine learning algorithms work.
Also learn python basics and look for appropriate APIs and look for the state of art of the subject.</td><td colspan="2">Planned start date: 01/3/2019
Planned end date: 01/06/2019</td></tr>
<tr><td colspan="2">Start event: 01/03/2019
End event: 01/06/2019</td></tr>
<tr><td>Internal task T1: Python basics and APIs.
Internal task T2: Look for state of art.
Internal task T3: NLP over structured text.
Internal task T4: ML algorithms (SVM) and neuronal networks.</td><td>Deliverables:
None</td><td>Dates:
01/06/2019</td></tr>
</table>

<table>
<tr><td>Project: Code Structure</td><td colspan="2">WP ref: 2</td></tr>
<tr><td>Major constituent: SW</td><td colspan="2">Sheet 2 of 5</td></tr>
<tr><td rowspan="2">Short description:
Start the ‘code skeleton’ and divide it by block/modules for an easier access and version control.
The code is not supposed to give any result yet in any stage.</td><td colspan="2">Planned start date: 10/4/2019
Planned end date: 20/06/2019</td></tr>
<tr><td colspan="2">Start event: 10/04/2019
End event: 20/06/2019</td></tr>
<tr><td>Internal task T1: Code flexible methods for the preprocess block.
Internal task T2: Code flexible methods for the learning module.
Internal task T3: Code the data visualization and validation module.</td><td>Deliverables:
Code update
Code update</td><td>Dates:
05/05/2019
13/06/2019</td></tr>
</table>

Project: Develop preprocess module	WP ref: 3	
Major constituent: SW	Sheet 3 of 5	
Short description: Code one or more preprocess methods using NLP libraries. This module will have text as input and a numerical vector as output.	Planned start date: 15/5/2019 Planned end date: 02/08/2019	
	Start event: 15/05/2019 End event: 02/08/2019	
Internal task T1: Code the bag of words Internal task T2: Implement tdfidf and other methods (also variations of each).	Deliverables: Code update Code update	Dates: 05/05/2019 02/08/2019

Project: Develop learning model module	WP ref: 4	
Major constituent: SW	Sheet 4 of 5	
Short description: Code one or more ML methods using python libraries. This module will have a numerical vector as input and the probabilistic results as output.	Planned start date: 23/6/2019 Planned end date: 02/09/2019	
	Start event: 23/06/2019 End event: 02/09/2019	
Internal task T1: Code the SVM method. Internal task T2: Code a custom CV.	Deliverables: Code update Code update	Dates: 02/08/2019 02/09/2019

Project: Statistics/Graph	WP ref: 5	
Major constituent: SW	Sheet 5 of 5	
Short description: Use the results information to visualize the statistics from each method using python libraries.	Planned start date: 14/6/2019 Planned end date: 23/07/2019	
	Start event: 14/06/2019 End event: 23/07/2019	
Internal task T1: Being able to visualize data results from validation module.	Deliverables: Code update	Dates: 02/08/2019

Milestones

WP#	Task#	Short title	Milestone / deliverable	Date (month)
1	1	Python basics and APIs	1/1	1
1	2	State of art	1/1	2
1	3	NLP	1/1	2
1	4	ML	1/1	3
2	1	Preprocess block	1/1	2
2	2	Learning block	1/1	2
2	3	Data visualization block	1/2	4
3	1	BoW	1/1	3
3	2	TDFIDF	2/3	4
4	1	SVM	2/3	4
4	2	Cross-Validation	2/4	5
5	1	Data visualization	2/3	4

2. Estat de l'art de la tecnologia utilitzada o aplicada al projecte:

2.1. Introducció a la intel·ligència artificial i a l'aprenentatge automàtic

La intel·ligència artificial és una branca de la ciència que lluita per una de les aspiracions més desafiants per la humanitat: aconseguir que una màquina sigui capaç d'imitar un comportament intel·ligent. Amb aquest objectiu han nascut una multitud de tècniques al llarg dels anys que han intentat aproximar-se tot el possible a aquesta fita. És per això que cal tenir clara la definició d'intel·ligència artificial, ja que hi pot haver varies interpretacions de fins on ha d'arribar aquesta intel·ligència per tal de ser considerada com a tal. Teòricament, i amb la definició sobre el paper, el simple fet de programar una màquina perquè realitzi amb una alta precisió una única tasca, com podria ser una màquina industrial programada per treballar un metall, ja correspon a un comportament intel·ligent, tot i que no se li pugui demanar cap altre tasca ja que no està preparada per dur-la a terme. Coneixent les seves limitacions, usem les màquines per realitzar tasques especifiques, perquè en aquestes, poden ser més ràpides que nosaltres. Hem de tenir en compte que per molt més potent que arribi a ser el nostre cervell, aquest està constantment realitzant tasques vitals i processant una enorme quantitat d'informació per mantenir un ordre i una salut al nostre cos. És per això que dipositem certa responsabilitat a les màquines per arribar a conclusions més exactes. Tot i així, fins al dia d'avui, encara no s'ha aconseguit que una màquina tingui la capacitat de realitzar múltiples tasques alhora de manera tant eficient (exceptuant alguns avenços recents dins el camp de la robòtica).

Ara, quan parlem d'aprenentatge automàtic, el context és lleugerament diferent. La intuïció ens porta a pensar que és quelcom relacionat amb que una màquina sigui capaç d'aprendre per sí sola, en el nostre context, reconèixer un patró entre les dades de les quals disposa. I bé, exactament d'això es tracta. L'aprenentatge automàtic centre els seus esforços en crear models matemàtics que proporcionin a una màquina la capacitat d'aprendre autònomament, és a dir, que un màquina imiti un comportament intel·ligent, però que a més aprengui a fer-ho sola. Ubicant aquesta definició i comparant-la amb la d'intel·ligència artificial, es pot apreciar que l'aprenentatge automàtic queda englobat en una branca dins el món de la intel·ligència artificial; tot i que avui en dia aparenti que només l'aprenentatge automàtic, a causa de la gran popularitat que ha guanyat durant les dues últimes dècades, formi part de la intel·ligència artificial.

Amb l'arribada de la digitalització, l'increment dels dispositius d'emmagatzematge i un canvi de mentalitat a l'hora d'apreciar el valor de les dades, s'ha entrat en una tendència d'acumular més i més informació. Això no ha fet més que beneficiar les capacitats de l'aprenentatge automàtic, ja que aquest, es nodreix de la gran quantitat de dades que ens envolten fins ser capaç de generar una predicció per registres d'informació que encara no ha processat mai.

Fora de les matemàtiques que donen sentit a tot això, un assumpte important a tenir en compte és aconseguir caracteritzar aquesta informació de tal manera que per cada mostra tinguem un seguit d'informació expressable en un vector, on cada valor d'aquest vector representa un atribut de la mostra. Per exemple, si la mostra és d'un pacient clínic,

un atribut pot ser el seu sexe, altura, edat o una mesura de les seves constants vitals. Amb una base de dades d'informació estructurada, es pot dur a terme una gestió d'aquestes dades per tal de veure quin comportament mostren, i d'aquí, proposar un model per fer prediccions per futures mostres o classificar-les en cas de ser mostres etiquetades.

Per tal d'optimitzar el seu rendiment, són necessaris també, en la gran majoria de casos, una gran quantitat d'exemples per tal de que el model sigui capaç d'aprendre amb precisió; sense deixar de banda la correcta selecció d'atributs a tenir en compte. És aquí on les empreses tecnològiques juguen el seu paper. I és que, si un observa els moviments de les empreses pioneres en el sector tecnològic, o les que poc a poc han pujat a l'onada, pot apreciar, com s'ha dit anteriorment, que la tendència ha estat acumular una gran quantitat de dades per així tenir una base d'informació que tractar i d'on poder extreure indicadors clau pels seus negocis. Aquesta tendència es podria pensar que és relativament recent, i de fet encara és un món creixent, però la tecnologia que hi ha al darrere, l'aprenentatge automàtic, ja portava més d'un segle coent-se.

2.2. Etapes inicials i creixement de la intel·ligència artificial

Un dels pioners en crear un model probabilístic que mirés d'adaptar-se a la realitat va ser Thomas Bayes l'any 1763 amb l'assaig *An Essay towards solving a Problem in the Doctrine of Chances*[3] (publicat dos anys després de la seva mort). Bayes en aquells temps ja va veure que existia una relació entre la realitat i la probabilitat, cosa que va demostrar per donar llum al conegut teorema de Bayes, basat en probabilitats condicionals i a priori. Aquest mètode va ser precursor dels models probabilístics bayesians i va ser mig segle més tard que es va expandir gràcies al esforços de Pierre-Simon Laplace per instaurar-lo i propagar-lo[4].

Just uns anys abans de que Laplace hagués publicat el seu treball donant forma al teorema de Bayes, va se quan un altra model va sorgir: els quadrats mínims. Aquest mètode va adquirir importància molt ràpidament, ja que figures tant rellevants com Gauss, el van emprar per preveure la posició d'astres[5], concretament es va usar per localitzar l'astre Ceres, i a més a més, va ser el mètode que el va ajudar a inventar la distribució normal. Aquest mètode s'utilitzaria més endavant per resoldre amb consistència problemes de regressió, problemes d'anàlisi de tendències que conformen una important porció dels problemes vigents.

Acostant-nos ja al segle passat, val la pena afegir que, a l'any 1913, Andrey Markov descriuria per primera vegada la tècnica que hauria utilitat per analitzar un poema: les cadenes de Markov[6]. Aquesta tècnica es fonamenta en definir estats dins un sistema i associar-los mitjançant la probabilitat que d'un estat es passi a un altre; establint així una relació entre estats. L'únic inconvenient d'aquest mètode és que només depèn de l'estat en que es trobi el sistema, independentment dels estats previs que hagin ocorregut, per tant, no té memòria, cosa que provoca que no s'adeqüi al plantejament del problema en alguns casos. Tot i així, per fer notori el seu pes dins el món tecnològic, és necessari realçar que aquest és un mètode tant potent que s'ha usat en projectes d'escala globals,

com ho és per exemple, PageRank de Google, per fer-se una idea de la magnitud de les capacitats que presenta.

Però no seria fins l'any 1951 que hi hauria un primer intent de recrear una xarxa neuronal. L'autor, Marvin Minsky, es va basar en la teoria Hebbiana, la qual proposa que l'eficiència de la transmissió d'informació entre neurones està relacionada amb l'activació simultània d'aquestes al reaccionar amb un mateix patró. És a dir, que si una neurona o conjunt de neurones s'activen sempre davant un mateix estímul, tendiran a activar-se les unes a les altres i a facilitar el trànsit d'informació entre elles. Així doncs, Minsky va ajuntar 40 maquines que simulaven aquest comportament i va aconseguir que aquestes guardessin uns pesos d'activació, simulant per primera vegada el funcionament del cervell[7].

Un lustre més tard, a mans de Frank Rosenblatt, sorgiria el Perceptron[8], que és equivalent al que entenem com a neurona dins una xarxa neuronal actual. Aquest invent seria el precursor de la tecnologia més explotada avui dia dins el sector. En poques paraules, cada Perceptron consisteix d'una regressió lineal dels valors d'entrada i comprovar si el valor resultant supera el llindar d'una funció d'activació, la qual com el seu nom indica, farà que la neurona s'activi i traspassi informació a la sortida[9]. Aquest descobriment va crear un gran rebombori mediàtic, generant altes expectatives que el van conduir a ser investigat en profunditat. Tot i les alegries del moment, Minsky juntament amb Seymour Papert, van discutir sobre les seves limitacions en el llibre Perceptrons[10], publicat l'any 1969. Una de les més importants era la complexitat computacional de les operacions d'algunes prediccions, com ho era la funció XOR. A més a més, exposa que inclús hi ha algunes prediccions que inclús als humans ens costa processar, reflectint-ho directament en la portada del llibre amb un contrast de colors i figures geomètriques expressament escollits per generar un efecte òptic (veure Fig. 1).

Durant la segona meitat del segle XX va sorgir el mètode de diferenciació automàtica[11] (o Backpropagation), que permet optimitzar la funció de cost de xarxes connectades fent un recorregut invers de d'aquestes. Va ser un descobriment que jugaria un paper d'importància per la consistència de les xarxes neuronals, però tot i així, pel que fa a implementacions de gran escala, els investigadors es trobaven amb limitacions de hardware, per la qual cosa l'evolució d'aquesta branca era constantment frenada tot i les seves ganes de créixer.

Coetàniament a aquesta època, van anar sorgint nous mètodes que intentaven enfocar els problemes des d'una altre punt de vista al de les xarxes neuronals. Un dels més coneguts és el Neares Neighbor, el qual es basa en decidir segons l'element o elements coneguts més pròxims al vector d'entrada[12]. Poc després es va inventar l'aprenentatge reforçat (o Reinforcement Leanring) amb la tècnica Q-learning[13], la qual és capaç d'optimitzar la predicció tenint en compte la recompensa del sistema en una cadena de Markov finita. Aquest concepte de recompensa influeix en la decisió final del mètode, en tant que si s'ocasiona un error serà penalitzat i si, en canvi, s'encerta, serà recompensat. Aqueta idea va ser pionera i s'implementaria també en altres tècniques que s'inclourien al món de l'aprenentatge reforçat. Finalment una altre a destacar és el conegut com Support Vector Machine[14], el qual està inclòs dins l'aprenentatge supervisat[15], ja que requereix d'etiquetes que determinin la classe de la mostra per aprendre. El que fa internament el mètode és generar una frontera entre les classes de tal manera que la

distancia entre la frontera i la distancia d'aquesta als punts que separa a una banda i altra sigui màxima.

Poc a poc la tecnologia va anar avançant i els impediments computacionals es van anar solucionant. Es van anar ajuntant xarxes neuronals especialitzades en certs tipus de decisió per tal de crear el que es coneix com *clústers*, d'aquesta manera la xarxa podia adquirir un enteniment més ampli sobre les dades i s'aniria donant forma al que avui en dia coneixem com aprenentatge profund. Aquest camp requereix d'immenses bases de dades i les màquines més modernes, computacionalment parlant, per tal d'entrenar i processar tota la informació, però, gràcies al seu abast, és el que les grans empreses utilitzen per obtenir els millors resultats. És per això que s'han utilitzat en projectes de gran escala i han estat els que més impacte han causat en els medis aquests últims anys. I és que els resultats obtinguts amb aquests mètodes han estat sorprenents pel món de la intel·ligència artificial.

2.3. Implementacions a gran escala

Les implementacions a gran escala són els millor exemples de triomfs en l'aprenentatge automàtic, ja que han estat projectes que han atret grans inversors i han involucrat als enginyers de les empreses pioneres del sector tecnològic. Empreses com Google, ofereixen serveis que usen algunes de les tècniques de gran escala, per exemple, el seu famós algoritme de ranking de pàgines web en seu buscador, PageRank, el qual s'ha comentat anteriorment. Netflix, per la seva banda, va proposar un concurs amb l'objectiu de millorar el seu sistema de recomanació per almenys un 10%[16]. Aquesta competició va estar disputada durant més de tres anys i finalment va haver-hi un guanyador que va integrar el seu algoritme al de la plataforma de reproducció de sèries i pel·lícules.

Una aplicació també molt estesa ha estat pel reconeixement facial. Aquesta, ha estat una aplicació inclosa en la funcionalitat de la majoria de dispositius mòbils actuals, usada sobretot en passos d'autorització. Ara bé, per tal d'entrenar aquests models, han estat necessaris molts exemples de mapes facials que, en aquest cas, poden no ser tant complicats d'aconseguir, però no és així per altres casos. Un exemple en són l'equip de Google Research, els quals volien entrenar un model perquè generés mapes de profunditat en imatges. L'equip tenia la infraestructura suficient per dur a terme el projecte, però els hi faltaven exemples suficients com perquè l'entrenament fos eficient. Va ser en aquest punt quan van tenir la brillant idea d'utilitzar un 'hashtag' de l'aplicació d'Instragram per localitzar una sèrie de vídeos, on els participant que apareixien en aquest estaven en posicions estàtiques durant el transcurs del vídeo mentre l'observador (el càmera) es desplaçava al voltant dels participants. D'aquesta manera van aconseguir un mapa de l'espai amb objectes estàtics que els va permetre entrenar el seu model. Aquest fet anecdòtic serveix per argumentar que per molt bones que siguin les tècniques i els algoritmes que s'utilitzen, de vegades l'enginyeria més rellevant pel problema és el mètode d'obtenció d'exemples i el tractament de les dades que fem d'aquests exemples.

Seguint amb altres implementacions a gran escala, una branca que ha obtingut bons resultats ha estat la de l'aprenentatge reforçat, el qual s'ha utilitzat en varis jocs per tal d'aconseguir que la màquina sigui capaç de jugar sola, i no només això sinó que sigui

capaç de superar els rècords mundials o guanyar a campions del món. El primer intent va ser amb el joc de les 'dames' i el 'tres en ratlla', a les dècades dels anys 50 i 60. Posteriorment, es fa fer un intent amb el 'blackgammon', però en cap d'aquests tres casos l'algoritme podia competir contra els millors del món en els jocs. No va ser fins la dècada actual (segona del segle XXI), que es va aconseguir guanyar a professionals. Primer va ser en un joc televisiu anomenat *Joepardy*[17], per part de l'IBM, i més recentment, amb una programa desenvolupat per Google, *AlphaZero*[18], el qual es centre en els jocs dels escacs, el shogi i el go.

Una de les tècniques utilitzades per guanyar en el joc de *Jeopardy*, l'IBM va utilizar tècniques de *Natural Language Processing*. Aquest camp tracta les dades en format de text i està format per un conjunt de tècniques amb les quals s'intenta extreure informació del text o analitzar-lo sintàcticament per trobar-hi el sentit. Les aplicacions que més han ressaltat al llarg del temps han estat utilitzades per detectar l'emoció que conté el text del missatge o per entendre les paraules utilitzades amb la finalitat d'establir una conversació. Un clar exemple de la seva utilitat és a l'hora de traduir textos, on cal una comprensió de l'estructura del text i el significat de les paraules en l'idioma escrit per tal de fer una traducció acerada.

Totes aquestes implementacions han estat més pròximes a l'ús quotidià de les persones, però n'hi ha d'altres en camps més específics, com en el camp de la medicina, on també s'ha contribuït en la millora de la presa de decisions tant importants com ho són el diagnòstic o detecció de malalties, punt en el que entrarem més en detall tot seguit.

2.4. Impacte sobre la medicina i la biotecnologia

Un dels camps que més s'ha beneficiat de l'aprenentatge automàtic ha estat el camp de la medicina. És poca tota contribució que es faci en aquest camp, ja que ajuda a mantenir la salut de la població, i la seva eficiència és clau per assegurar l'esperança de vida d'un pacient. En aquest sentit un munt d'investigadors van veure l'oportunitat d'utilitzar les dades clíniques dels pacients per tal d'elaborar models que estimessin si els candidats eren portadors d'una malaltia o no.

D'aquesta iniciativa van formar-se centres d'informació amb la motivació d'emmagatzemar i analitzar aquesta informació. Un clar exemple és el NCBI, que com ja s'ha comentat, integra diverses bases de dades amb literatura mèdica, seqüencies de gens, proteïnes i genomes, variacions heretables d'ADN i estructures moleculars.

Amb bases de dades similars, s'han fet estudis per detectar patologies cancerígenes en imatges, com van fer a l'estudi *Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening*[19]. Els resultats d'aquests investigadors van mostrar que las decisions preses per una xarxa neuronal convolucional sobre més d'un milió d'imatges va ser millors que els predits pels radiòlegs, un fet sorprenent i de gran importància donada la conseqüència d'encertar en els diagnòstics.

Així doncs, amb aquests encoratjadors resultats, tot apunta a que la intel·ligència artificial acabarà jugant un paper clau en un futur no molt llunyà, on els metges i investigadors de

patologies podran recolzar-se en aquestes tecnologies de manera habitual per afiançar millors resultats en els diagnòstics.

2.5. Natural Languaje Processing

Tornant enrere en l'aprofundiment del punt anterior, dins del món de la intel·ligència artificial cal destacar la utilitat que el processat del llenguatge natural ens aporta. Aquesta branca és la dedicada a estudiar el tractament del llenguatge, ja sigui oral o escrit, i es pot dividir en dos components: la comprensió del llenguatge i la generació del llenguatge. La primera de les dues és la que acostuma a ser més complicada, ja que depen de que el sistema sigui capaç d'extreure'n la correcta informació, mentre que la segona depen més de la interpretació interna que usa el sistema per generar el text. Per tal d'aconseguir extreure informació del text, tenim varies tècniques que podrem usar en funció de la utilitat que aportin. Aquestes tècniques són les usades per preprocessar el text i així amb posterioritat passar més informació al model amb el que estiguem treballant (normalment un model clasificador).

Per començar, el més comú es *tokenitzar* el text. Un *token* és una unitat dins el text, és a dir, pot ser una paraula, un signe o qualsevol element que confomi el text. D'aquesta manera tindrem separades totes les paraules del text i podrem tractar-les en detall.

Per tal de simplificar els text i veure més fàcilment de quines paraules està format, s'utilitzen dues tècniques que agafen els *tokens* i els manipulen. La primera i més senzilla és anomenada *steming* i consisteix en treure prefixos i sufixos a les paraules i transformar-les en la seva versió original. Ara bé, hi haurà casos on no funcioni correctament, i és per això que també s'utilitza el *lemmatization*, el qual consisteix en transformar la paraula en el seu *lemma*, però fent ús d'un diccionari. S'enten com a *lemma* la paraula d'orígen o arrel de la paraula *tokenitzada*; un exemple en la llengua anglesa és en el cas de les paraules *gone, going* i *went*, les quals seran totes transformades a la paraula *go*, però no seria el mateix resultat en el cas d'haver utilitzat *stemming*.

Després de tot això es pot afegir a les paraules un tag segons si són verbs, adjectius, noms, etcètera i d'aquesta manera ja tenir una primera estructura de la composició del text. Tot i així hi ha la limitació de que l'algoritme desconeix els noms propis, és per això que existeix el *entity recognition*, que analitza la resta de la frase per intentar desxifrar si el nom propi és un lloc, una institució, una persona, una quantitat o d'altres.

Finalment, un cop tenim el text descompost i etiquetat, podem passar a ajuntar tota aquesta informació per analitzar sintàcticament la frase. D'aquesta manera obtindrem al final una estructura de frase amb els subjectes i predicats.

Tota aquesta manipulació del text ha servit per desenvolupar projectes d'anàlisi del sentiment en el text, com ho fa per exemple Facebook. També tenim el cas en expansió dels chatbots o assistents virtuals, els quals també fan ús d'aquests processos per tal d'entendre i generar respostes. Dins l'assistència automàtica també tenim els innovadors casos de Windows, Amazon o Apple, per exemple, amb els seus corresponents assistents Cortana, Alexa i Siri els quals utilitzen el reconeixement de la parla per

transformar les paraules a text i posteriorment analitzar-lo i processar el que els usuaris desitgen.

També és aplicable en traduccions de text com ho ha fet Google amb el seu traductor, el qual recentment ha fusionat varies tecnologies per traduir el text en temps real usant la càmera dels dispositius mòbils.

A més a més, també té aplicacions per extreure les paraules clau que puguin aparèixer al text o extreure certa informació, sobretot quan se'n sap l'estructura.

Aquests últims no han estat tant estesos com els altres degut a que la seva utilitat no ha presentat, de moment almenys, una escalabilitat tant alta com ho estan fent les seves aplicacions germanes. Tot i així, en aquest projecte són de les que es faran servir ja que s'adeqüen al propòsit que més endavant s'exposarà en detall.

2.6. Extracció de característiques de text

La motivació de l'extracció de característiques és aconseguir dades que identifiquin el text que s'analitza i poder-lo treballar des d'un altre punt de vista que no sigui la mera lectura, ja que els models d'aprenentatge automàtic requereixen de mètriques numèriques o etiquetes finites per treballar.

Amb les tècniques de *Natural Language Processing*, podem extreure bastantes característiques del text, algunes d'elles són el nombre de paraules que conté el text o la longitud mitja de les paraules utilitzades. Ambdues característiques ens poden ajudar a distinguir entre una conversació rápida entre dues persones o una redacció més elaborada sobre un tema, per exemple.

Altres utilitats calculen el ratio de sentiment del text per analitar les opinions globals de la gent en, per exemple, un producte, però també podria ser una empresa o una persona. És a dir que també es poden analitzar missatges de xarxes socials, com s'acostuma a fer amb Twiter, per detectar possibles casos de discriminació o similars.

Finalment, dins aquest apartat és imperatiu anomenar les tècniques *bag of words* i *time frequency – inverse document frequency*. Ambdues necessiten varis documents com a valor d'entrada, i el que fan és transformar les paraules en valors. Per fer-ho, les dues tècniques fan ús d'un mètode similar on la principal diferència és que el *bag of words* només es centre en el nombre de vegades que una paraula apareix en el corpus, mentre que el *tf-idf* té en compte també la freqüència amb la que la paraula apareix en el donat llistat de documents, en altres paraules, afageix un component de reputació a les paraules a l'hora d'evaluar-les.

Totes aquestes mètriques extraïbles són molt útils en el cas de voler extreure dades especifiques del text, ja que genera una caracteritazció per les paraules de tal manera que les usades de manera similar tindran un vector d'informació, que si es representa en les seves dimensions, mostraran una proximitat amb elevada probabilitat. Un exemple de la seva utilitat per aquest casos es descriu en els articles referenciats aquí[20][21].

3. Metodologia/desenvolupament del projecte:

Abans d'iniciar amb el desenvolupament del projecte, recordar al lector que l'objectiu principal és detectar en un conjunt d'articles mèdics quantes persones han estat sota estudi o quantes mostres s'han tingut en compte a l'escriure l'article. Per esclarir el concepte posarem un petit exemple.

Si l'artícle conté una frase com la següent:

"Aquest estudi s'ha realitzat sobre 3500 dones Egípcies entre 45 i 65 anys l'any 2016."

L'objectiu del projecte és detectar el nombre 3500 i descartar els nombres 45, 65 i 2016, ja que aquests expliquen edats i dates i, per tant, no són objectiu.

En aquest exemple el candidat (número) es tracte de les dónes de les qual s'han tret mostres, però l'article ens podria parlar de pacients, mostres, casos o qualsevol altra subjecte a ser sotmès a estudi.

Havent esclarit el problema a resoldre, s'iniciarà la descripció del seu desenvolupament.

3.1. Observació d'informació inicial i descàrrega de dades

Gràcies a la col·laboració amb l'ICO, el supervisor ens va proporcionar una arxiu Excel amb la informació necessària per poder iniciar el projecte (veure Taula 1). Com es pot apreciar a la taula, aquest Excel conté varies columnes, entre elles, les més útils ens són l'identificador d'article i la columna on apareixen els candidats objectiu, ja que gràcies a la primera, podrem accedir als articles per identificador i la segona la usarem més endavant per construir les etiquetes que diferenciaran la classe. Això és important ja que com que el problema és un problema de classificació binaria (si és o no és objectiu) els models que millor s'ajusten són els d'aprenentatge supervisat. Aquests necessiten d'una etiqueta que distingeixi la classe a la que pertany cada mostra durant l'entrenement, per així utilitzar-les per supervisar la seva optimització.

Amb la columna d'identificadors doncs, podrem accedir a ells amb l'ajuda de la llibreria Bio, i dins d'aquesta, del mòdul Entrez. La llibreria Bio compren un conjunt de mòduls que faciliten l'accessibilitat a les bases de dades de la NCIB mitjançant la seva API REST, amb la qual es poden demanar diferents recursos a través d'HTTP. En el nostre cas, com ja s'ha comentat, s'usarà el mòdul Entrez per accedir a Pubmed, la base de dades centrada en guardar abstractes d'articles mèdics.

S'ha escollit tractar els abstractes enlloc dels articles sencers ja que en la gran majoria d'aquests, el nombre de pacients sotmesos a les proves ja hi queda reflectit. A més a més, es produiria una descompensació massa gran entre els casos positius i negatius (entenent com a cas positiu el nombre objectiu i com a cas negatiu qualsevol nombre no objectiu), així com un increment en el temps computacional al haver de processar tots els articles sencers.

Així doncs, procedirem a descarregar els abstractes pels que tinguem identificadors a l'arxiu Excel i generarem un nou dataframe amb la nova informació extreta (veure Taula 2). Amb la idea de tenir quanta més informacions millor, aquest dataframe no només

inclou l'abstracte sinó que també s'hi ha inclòs autors, dates de publicació, el llenguatge i d'altres.

Una pràctica que es seguirà durant el projecte és, un cop generat el dataframe, guardar-lo en un directori del projecte en format pickle. D'aquesta manera evitarem haver de tornar a executar parts de codi innecessàriament.

3.2. Extracció de dades en format text

Amb aquest dataframe creat, podrem passar a preprocessar els abstractes per tal d'extreure característiques d'aquests i generar un nou dataframe on hi haurà una fila per cada nombre detectat als abstractes. D'aquesta manera tindrem tot el repertori de candidats en un mateix dataframe i podrem anar afegint informació per caracteritzar-los pas a pas.

Primer de tot i de les coses més importants a l'hora de treballar en aprenentatge automàtic és separar bé les mostres que utilitzarem d'entrenament i les que utilitzarem per testejar. A l'inici del projecte vam cometre l'error de separar el dataframe després d'haver extret informació i això va corrompre les dades, fent que els resultats fossin sorprenentment bons (comentats més endavant).

A més a més, per afegir consistència als resultats, s'ha programat una cross-validation que té en compte tant la fase de preprocessat com la de classificació. Per dur-ho a terme s'han barrejat les mostres aleatòriament per tot seguit dividir-les en 10 blocs i formar els dos dataframes (1 bloc per test i 9 blocs per entrenament). El procés es repeteix 10 vegades, cadascuna iterant el bloc de test i agafant la resta de blocs per entrenar. Amb aquest mètode de treball s'han fet varies proves barrejant el dataframe original de diferents maneres per veure si els resultats són similars i donar robustesa al model.

Seguint amb el desenvolupament de l'extracció de dades, per començar a analitzar el text haurem de fer ús de la llibreria NLTK de Python. Com ja s'ha introduït amb anterioritat, aquesta llibreria ens permet manipular els textos per extreure'n dades mitjançant tècniques de NLP. Al llarg del projecte s'han anat introduint funcionalitats a aquest bloc en tant que s'ha anat millorant la profunditat de l'anàlisi i cada vegada s'ha extret més informació.

En un principi vam començar *tokenitzant* cada abstracte, d'aquesta manera es separen totes les paraules dels textos en unitats, normalment paraules o signes de puntuació. Però més endavant ens vam adonar que necessitàvem netejar alguns tokens, generalitzar-los i afegir-hi algun sentit més per afinar els resultats.

Amb aquest objectiu, tots els abstractes són descompostos en frases. Una de les funcions de la llibreria NLTK ens permet fer-ho en una línia de codi. Un cop dividits en frases, passarem a analitzar-les en més profunditat. Primer cal *tokenitar* la frase per obtenir les unitats bàsiques amb les que la llibreria treballa, després es passarà a *lemmatitzar* les paraules que hi apareguin per minimitzar el nombre de derivacions possibles de les paraules. S'ha decidit usar la *lemmatització* per davant del *stemming* perquè, recordem, el *stemming* freqüentment provoca que paraules derivades no es

transformin en la seva paraula arrel, sinó en una altra paraula. Un exemple esclaridor és el cas següent en la llengua anglesa:

- Streming: Caring – Car
- Lemmatization: Caring – Care

Ara bé, per tal de *lemmatitzar* correctament, és necessari etiquetar les paraules segons la seva funció dins la frase, és dir, si són noms, verbs, adjectius, adverbis o algun altre, ja que una paraula pot tenir varis *lemmas* depenent del context. D'aquesta manera si es proporciona aquesta etiqueta, es pot resoldre aquesta ambigüitat i *lemmatitzar* correctament en més casos.

Un cop obtinguem les paraules amb els seus *lemmas* i les seves etiquetes, haurem de netejar de signes de puntuació i d'altres signes les frases per evitar que siguin entesos com a paraules, i ara sí, podrem passar al *chunking*. El *chunking* segueix una gramàtica concreta que nosaltres mateixos hem elaborat per aquest projecte i té com a objectiu assegurar que hi hagi el màxim nombre de candidats inclosos dins un sintagma nominal el més complert possible. Amb aquesta pràctica tindrem a la nostra disposició tots els sintagmes nominals del candidats, sintagmes que usarem com a context del candidat. Seguint en aquesta via, procedirem a revisar frase per frase els *tokens* que la componen i revisarem si són números. La revisió descarta directament valors inferiors a 2 i els nombres decimals, ja que segur que no són representatius.

Un cop trobat un candidat en el text, l'afegirem al dataframe i ens quedarem amb la frase sencera on es troba, les paraules del seu sintagma nominal amb la distancia d'aquestes al candidat, i si el nombre apareix en forma de fracció (ja que hi ha vegades que el candidat està expressat només d'aquesta manera, havent-nos de fixar només en el denominador de la fracció).

Amb aquesta informació i afegint si el candidat és de la classe positiva o negativa, fent ús de la informació del dataframe original (on teníem els casos positius llistats per cada article), ja haurem acabat amb la primera fase de l'extracció de dades.

La raó de fer tot aquest procés és per utilitzar un mètode que converteix les paraules de text en pesos, l'anomenat TF-IDF, que usa aquesta funció per calcular els pesos:

$$\boldsymbol{TFIDF(w) = TF(w) \cdot IDF(w) = TF(w) \cdot log(\frac{|D|}{\{D: w\ in\ D\}})}$$

Aquest mètode també té dues fases: la d'entrenament, on donats uns textos s'aprèn un vocabulari i es calculen tots els pesos d'aquest seguint la fórmula anterior; i la de transformació, la qual transforma un text d'entrada tenint en compte el vocabulari i els pesos apresos durant l'entrenament.

Nosaltres, tot i que no es va fer des d'un inici, farem servir només els contextos, és a dir, els sintagmes nominals, dels candidats que sabem que són positius i els farem servir per entrenar el model del TFIDF (només els del dataframe d'entrenament per cada iteració).

Utilitzant després el model entrenat per transformar tots els contextos dels candidats, obtindrem, per cada candidat, una llista amb el vocabulari del TFIDF i uns pesos per

cada paraula en funció del context que li haguem passat. Si, per exemple, no apareix una paraula del vocabulari en el context a transformar, el valor d'aquesta paraula serà zero, sinó tindrà un valor superior.

Ara només faltarà afegir aquesta informació als dataframes de test i entrenament i disposarem de la informació suficient com per passar a classificar els candidats (veure Taula 3).

3.3. Classificació dels candidats

Aquest bloc de treball, com el títol il·lustra, està construït per tractar els algoritmes de classificació utilitzats al llarg del projecte. A causa d'algunes incidències, no ens ha donat temps de fer més proves amb més d'un classificador, tot i que el que hem utilitzat és molt flexible.

El model classificador que hem analitzat ha estat el SVM, el qual consisteix en trobar una frontera de decisió que tingui un marge màxim entre els punts que estiguin més a prop de la frontera a banda i banda (veure Fig. 2). Se'l considera molt flexible perquè no només és capaç de resoldre problemes de classificació lineal, també pot fer-ho per classificacions no lineals. El que utilitza SVM per aconseguir aquest resultat és el que es coneix com a kernelització. La kernelització el que fa és projectar les característiques en un espai amb més dimensions de forma que les dades siguin linealment separables. Llavors el model troba una frontera i es calcula la funció inversa de la projecció per tornar al conjunt de dades original (veure Fig. 3).

No només això sinó que ens permet ajustar, pels casos d'usar kernel, un hiperparàmetre que controlarà l'amplada i altura de les corbes usades per trobar fronteres: el paràmetre gamma. Tant amb alts com amb baixos valors d'aquest paràmetre es tendirà a sobreajustar la frontera. Aquest comportament del paràmetre ve donat per la seva relació amb la distribució normal (veure Fig. 4). I finalment, també ens deixa controlar el pes de la penalització per error de classificació: el paràmetre C. Aquest paràmetre ens permetrà ser més permissius o més estrictes amb els errors comesos, de tal manera que la frontera quedarà més o menys ajustada (veure Fig. 5).

Així doncs, amb la llibreria SKLEARN implementarem aquest mètode i amb els dataframes de test i entrenament per cada iteració, passarem els atributs dels candidats (pesos i distancies de paraules del context i un atribut binari sobre si és un nombre que ve d'una fracció) al model. Primer de tot el farem entrenar amb les dades d'entrenament i després farem una prova amb les dades de test. A l'haver partit en 10 blocs i haver usat un d'ells com a test per a cada iteració del procés, a l'obtenir resultats d'un bloc podem reconstruir el dataframe sencer de nou i analitzar els resultats totals.

El mètode l'hem provat varies vegades, algunes vegades amb més atributs que d'altres i també variant els pesos del paràmetre d'error C. Per exemple, amb aquest paràmetre hem pogut compensar el desequilibri de la proporció de classes, ja que hem incorporat un pes per C més estricte al cometre un error al classificar malament una mostra positiva que no pas en direcció contraria.

L'atribut gamma, en canvi, l'hem mantingut constant ja que el mètode permet atribuir-li el valor 'scale' que directament calcula el valor de gamma tenint en compte la desviació de les dades d'entrada, i per tant, al ser un paràmetre que prové de la distribució normal, s'assimilarà a una gaussiana amb variància equivalent a la de les dades d'entrada.

Aquesta part del codi ens retornarà tres vectors de la dimensió del dataframe complet: el primer amb la classe real a la que pertanyen les mostres, el segon amb les prediccions del classificador i el tercer amb la probabilitat de que la mostra sigui de la classe positiva.

Amb aquests tres vectors en tindrem suficient per analitzar els resultats a l'últim bloc del codi, dedicat al càlcul de mètriques i visualització d'aquestes.

3.4. Mètriques i visualització de dades

Com que el següent apartat està dedicat en la seva totalitat a exposar els resultats, en aquest apartat ens dedicarem a introduir i explicar quines mètriques hem usat per orientar-nos.

El més comú és representar la matriu de confusió abans de res, ja que d'aquesta en podrem extreure molta informació. La matriu, al ser el nostre un cas binari, serà una matriu 2x2 on les files estaran representades per les prediccions del classificador i les columnes per la realitat. D'aquesta manera ens quedarà la següent representació (veure Fig. 6), on a cada posició hi tindrem:

- (0,0) - # mostres predites com a positives i que són positives – *True Positives*
- (0,1) - # mostres predites com a positives i que són negatives – *False Positives*
- (1,0) - # mostres predites com a negatives i que són positives – *False Negatives*
- (1,1) - # mostres predites com a negatives i que són negatives – *True Negative*

D'aquesta matriu en podem treure totes les probabilitats i estadístiques necessàries per l'anàlisi dels resultats, com la probabilitat d'encert total o, per un millor estudi, el que es coneix com sensibilitat, precisió o també el False Positive Rate. La sensibilitat avalua si hem estat capaços de detectat la classe positiva, la precisió avalua si els que hem classificat com a classe positiva realment ho és i el False Positive Rate indica percentualment si hem classificat malament la classe negativa. Seguint aquest criteri es pot deduir que:

- $\text{Sensibilitat} = \frac{Ture\ Positive}{True\ Positive + False\ Negative}$
- $Precisió = \frac{True\ Positive}{True\ Positive + False\ Positive}$
- $FPR = \frac{False\ Positive}{False\ Positive + True\ Negative}$

Aquests tres valors ens seran útils per calcular altres mètriques més complexes. Una d'elles és l'anomenada F-Socre, que no és res més que la mitjana harmònica de la sensibilitat i la precisió, de tal manera que si ambdues tenen el valor màxim, el resultat serà el màxim, però si una d'elles és zero, el resultat serà zero.

- $F - Score = 2 \cdot \frac{Sensibilitat \cdot Precisió}{Sensibilitat + Precisió}$

Així podrem controlar que el llindar de decisió no s'estigui decantant massa per resoldre correctament una de les classes, ja que si la sensibilitat és molt alta, en la majoria dels casos, la precisió baixarà i viceversa.

Una altra mètrica molt interessant és el coeficient de Kappa Cohen[22].

- $k = \frac{Po - Pe}{1 - Pe}$

On P_o és la probabilitat d'acord, és a dir, el percentatge d'encert, i P_e és la probabilitat d'encert per atzar. Aquesta segona probabilitat és la més característica del coeficient, ja que, el que estipula és que tot model tendeix a encertar per atzar quan té un dubte, i per això aquesta mètrica ens compara la relació entre el nostre model i un model perfecte tenint en compte la probabilitat d'encert per atzar. És com si ens féssim la pregunta: quina quantitat de nova informació estic aportant sabent que per atzar encertaré un donat percentatge de vegades?

En resum, ens indica si el nostre model aporta més que el simple atzar.

Per últim, usarem les probabilitats de que els candidats siguin de la classe positiva (vector extret del mòdul de classificació), per tal de generar dues gràfiques molt representatives: la ROC i la Sensibilitat-Precisió.

La gràcia d'aquestes corbes és que aprofiten aquesta probabilitat per generar varis llindars que canviïn els resultats, ja que el SVM només permet posar un llindar a 0,5. D'aquesta manera obtindrem diferents matrius de confusió i al mateix temps diferents valors de sensibilitat i precisió, cosa que ens permetrà obtenir un seguit de punts (on els eixos són aquestes dues variables) i analitzar la seva eficiència màxima.

Pel cas de la ROC[23], estarem comparant la sensibilitat envers el ràtio de falsos positius, és a dir, la quantitat de casos positius que hem detectat comparat amb el nombre de negatius que hem detectat malament. Això és així perquè normalment, si posem un llindar que detecti molts casos positius, és a dir, una sensibilitat alta, segurament és perquè hem ajustat tant el model que també estem incloent elements de la classe negativa a la classe positiva, i això farà que el model en surti perjudicat.

De manera molt similar funciona la corba Sensibilitat-Precisió[], però aquesta compara la sensibilitat amb la precisió al llarg de les iteracions del llindar. Amb aquesta segona gràfica ens podrem fixar més en el comportament de la classe positiva i deixar de banda ràtios que tenen en compte l'avaluació dels casos negatius. En casos com el del projecte, on la proporció entre classes està descompensada, acostuma a ser més representatiu observar aquesta gràfica que no pas la ROC.

Per últim i per tenir una idea de com es poden comparar diferents corbes, el que s'acostuma a fer és mirar l'àrea per sota la corba, i així tenir un valor numèric que ens indiqui si el resultat és millor o no que un altre.

4. Resultats

Inicialment es van fer proves amb un mòdul de preprocessat no només poc elaborat, sinó que directament corromput. És per això que com es pot veure a la Fig. 7, la corba ROC ens va sortir sorprenentment bé, perquè havíem estat utilitzant informació de tot el dataframe per crear el TFIDF, i al transformar les paraules, totes tenien un valor assegurat ja que s'havien utilitzat per entrenar-lo. Per fer-se una idea numèrica, l'àrea sota la corba va ser de 88.77%.

Òbviament, aquest resultat va fer saltar les alarmes i ràpidament es va fer millores per evitar que es corrompés la informació i els blocs de test i entrenament quedessin ben aïllats. Això es va aconseguir implementant una cross-validation pròpia explicada anteriorment en aquest informe més en detall.

Els resultats aquesta vegada van ser més realistes (veure Fig. 8 i 9), i de fet va ser el punt de partida des d'on es va començar a intentar incorporar millores i nous passos. Com es pot veure en les figures, es va assolir una bona fita pel que fa a la ROC i Sensibilitat-Precisió, cosa que suggeria que s'havia trobat una bona pràctica a l'hora de caracteritzar els candidats. En aquest punt el percentatge d'encert total va ser del 91% (tot i no ser massa significatiu degut a la descompensació).

Lluny de donar-nos per satisfets, però, es va millorar el mòdul de preprocessat per fer un anàlisi més a fons de les paraules i entendre-les millor, però ens vam adonar que a l'haver afegit les distancies de les paraules com atributs del candidat, va fer empitjorar lleugerament els resultats (veure Fig. 10 i 11). En aquell moment ens va quedar clar que per afegir molta informació amb poca rellevància més val no posar-la des d'un principi, ja que per la majoria de candidats prendrà el mateix valor.

Després d'uns pocs intents i d'incorporar l'ordre aleatori abans de partir el dataframe en 10 blocs, es va acabar d'incorporar també tota la funcionalitat descrita al bloc de preprocessat i fent una prova es van aconseguir les següents mètriques (veure Fig. 12 i 13):

- Encert total: 92.55%
- Sensibilitat: 51,5%
- F-Score: 0.5447368421052632
- Kappa Cohen: 0.5042937105411952
- AUC ROC: 0.8111022129916454
- AUC PR RE: 0.4955654940041628

Com es pot observar, s'ha aconseguit un alt percentatge d'encert, tot i que només el 50% dels casos positius s'han classificat correctament. La resta de paràmetres ens indiquen que es manté una estabilitat en el sistema i que seguim aportant una millora respecte la probabilitat d'encert per atzar. A més, per tal de veure si l'aleatorietat afectava a l'obtenció de millors resultats es van dur a terme varies proves posant una llavor diferent al generador de nombres aleatoris. Afortunadament la variació de resultats va ser mínima, fet que fa el model més robust.

Finalment es van fer retocs al classificador per tal d'ajustar la penalització per error i, d'aquesta manera decantar la frontera i equilibrar la descompensació de classes. Es va fer igual que en el cas anterior i es van fer varies proves amb diferents llavors per veure si els resultats eren similars, i ho van ser (veure Fig. 14 i 15). I no només això sinó que van ser els millors resultats del projecte, inclús comparables a les primeres etapes del projecte quan vam corrompre la informació, a diferencia de que aquesta vegada s'havia tingut en compte les bones pràctiques:

- Encert total: 91.68%
- Sensibilitat: 75%
- F-Score: 0.6124497991967872
- Kappa Cohen: 0.5678065036508746
- AUC ROC: 0.8571127264620293
- AUC PR RE: 0.5276850110011502

Si comparem aquests resultats amb els anterior, podem apreciar que l'encert total és menor, però tot i així, estem detectant molts més candidats positius; és a dir que en el cas anterior estàvem resolent millor la classificació però tenint en compte els casos negatius. Si només tenim en compte els casos positius, pràctica habitual en aquests casos de descompensació entre classes, podem veure que havent sacrificat poca precisió, hem augmentat gairebé un 25% la sensibilitat del model.

Aquests han estat els resultats finals del projecte i, es pot dir que s'ha assolit l'objectiu de detectar la majoria dels candidats positius mantenint un nivell prou constant de precisió, la qual ronda el 50%.

5. Presupost

Com que no s'ha utilitzat cap software, ni instal·lació, ni s'ha sol·licitat cap servei a tercers i ni tant sols s'ha plantejat en ninguna etapa del projecte, el pressupost es calcularà amb una aproximació de les hores empleades per realitzar el projecte.

Així doncs, observant un estudi de tendències del mercat elaborat per PageGroup, podem agafar la referencia d'uns 20€/h, però per ser més realistes es tindrà en compte com si fossin 14€/h.

Així doncs, si tenim en compte que el projecte va iniciar al voltant de Març, han passat 7 mesos de mitjana de 30 dies. D'aquests 210 dies, com que s'ha estat treballant i estudiant s'aproximarà que s'ha pogut treballar 1/3 d'aquests dies. Dels 70 dies que resten, han estat aprofitades 8h al dia, per tant, això ens deixaria amb 560 hores treballades aproximadament. Si ara multipliquem aquestes hores pels 14€/h que estimem, ens surt que el pressupost aproximat del projecte és de 7840€.

6. Conclusions i futur desenvolupament:

Per concloure amb l'exposició del projecte, s'exposaran les conclusions que s'ha extret al llarg del desenvolupament d'aquest.

Per començar m'agradaria recordar que aquest és un camp poc explorat i que té pocs referents amb els que trobar similituds d'implementació, és per això que no teníem clar, en un inici, quins serien els resultats i, al final, han superat les expectatives. Tot i així, per tal de confirmar la seva robustesa serien necessaris més articles dels quals minar la informació i així poder fer més avaluacions. En el món de l'aprenentatge automàtic és un factor clau el fet de tenir bases de dades extenses per tal d'ajustar els models amb exactitud i aconseguir que rendeixin a alts nivells.

També és fonamental entendre les maneres de fer del NLP per poder profunditzar en la temàtica i ser capa d'extreure'n suc, ja que si no es va amb cura, pot arribar a ser contraproduent. Per exemple en el projecte no s'ha fet ús del que es coneix com *stopwords*[], les quals són paraules que es descartaran del llistat de *tokens*, ja que, podria ser que ens haguèssin descartat paraules pròximes a candidats positius, i hagués estat información clau que podríem haver perdut. De totes maneres, hi ha practiques, com netejar el text de signes, que milloren el rendiment del NLP i s'han de conèixer per tenir-les en compte.

A més a més, hem pogut comprovar que és molt fácil caure en males practiques, sobretot quan no es té una experiencia o un recorregut en el sector. Hem pogut tastar de primera mà la importància de mantenir-se alerta amb cada incorporació o modificació de codi que es fa i també de tenir clar en tot moment per quina etapa passaran les dades i quina forma es desitja que tinguin.

Sense ànim de menyspreu, la majoria de tècniques pioneres fa l'efecte que només utilitzin xarxes neuronals per resoldre els problemes, i bé, no és exactament una mala pràctica, però lligant amb el que s'ha comentat més amunt (i per les xarxes neuronals especialment), no sempre són necessaries per aconseguir bons resultats, a més de que són les que s'han apropiat del del camp del Deep Learnig i, per entrenar xarxes amb tantes capes, calen moltes mostres.

Tot i usar les metriques per comparar els models i les variacions que san fet al codi, cal tenir clar que no existeixen uns valor òptims d'aquestes mètriques, el més important és anar provant fins trobar amb el llindar adequat pel problema en qüestió. En aquests casos s'ha de tenir en compte si interessa més acertar tots els candidats positius però classificar com a positius exemples negatius, o al revés, acertat menys casos positius a costa d'acertar-ne més de negatius. És un assumpte subjectiu i dut a discusió en molts articles i escrits.

Cara un futur pròxim, la idea es fer un desplegament web del model per tal d'incorporar un front-end més visual al projecte i a més poder ensenyar amb mes claredat la utilitat i aplicació del projecte cara als investigadors. La pàgina disposaria, almenys, d'una buscador d'article i una zona dedicada al display de l'abstracte i estadístiques de l'avaluació del model.

Una altra possible, i factible ruta a seguir desenvolupant, és aprofitar que al codi és fàcil d'incorporar-hi mètodes alternatius, apart d'una xarxa neuronal, es podria provar un random forest, posat que treballem amb bastantes variables, podria ser útil utilitzar-lo per la seva simplificació de variables a tenir en compte per cada arbre aleatori que genera.

En definitiva, i per sobre de tot, la conclusió a la que s'ha arribat és que s'han assumit una bona quantitat de nous conceptes, entre ells, s'ha après a programar en Python, a treballar amb APIs REST, a manipular dades i extreure'n nova informació, a utilitzar models d'aprenentatge automàtic de varis camps, el significat d'algunes de les mètriques més representatives i d'altres relacionades amb aquest món que s'espera seguir desenvolupant d'ara en endavant.

Bibliografia:

[1] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So and Jaewoo Kang. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". *Bioinformatics*, 2019. [Online] Available: https://arxiv.org/pdf/1901.08746v3.pdf. [Accessed: 26 June 2019].

[2] Quinten McNamara, Alejandro De La Vega, Tal Yarkoni. "Developing a comprehensive framework for multimodal feature extraction". University of Texas at Austin Austin, Texas, 20 Feb 2017 [Online] Available: https://arxiv.org/pdf/1702.06151v1.pdf. [Accessed: 26 June 2019].

[3] Bayes, Thomas (1 January 1763). "An Essay towards solving a Problem in the Doctrine of Chance" (PDF). *Philosophical Transactions*. 53: 370–418. doi:10.1098/rstl.1763.0053. JSTOR 105741. Retrieved 15 June 2016.

[4] O'Connor, J J; Robertson, E F. "Pierre-Simon Laplace". School of Mathematics and Statistics, University of St Andrews, Scotland. Retrieved 15 June 2016.

[5] Legendre, Adrien-Marie (1805). *Nouvelles méthodes pour la détermination des orbites des comètes* (in French). Paris: Firmin Didot. p. viii. Retrieved 13 June 2016.

[6] Hayes, Brian. "First Links in the Markov Chain". *American Scientist*. Sigma Xi, The Scientific Research Society. 101 (March–April 2013): 92. doi:10.1511/2013.101.1

[7] Crevier 1993, pp. 34–35 and Russell & Norvig 2003, p. 17.

[8] Rosenblatt, Frank (1958). "The perceptron: A probabilistic model for information storage and organization in the brain" (PDF). *Psychological Review*. 65 (6): 386–408. doi:10.1037/h0042519

[9] Cireşan, Dan Claudiu; Meier, Ueli; Gambardella, Luca María; Schmidhuber, Jürgen (21 de septiembre de 2010). «Deep, Big, Simple Neural Nets for Handwritten Digit Recognition». *Neural Computation* 22 (12): 3207-3220. ISSN 0899-7667. doi:10.1162/neco_a_00052.g

[10] Minsky, Marvin; Papert, Seymour (1969). *Perceptrons: An Introduction to Computational Geometry*. MIT Press. ISBN 0-262-63022-2.

[11] Rumelhart, David; Hinton, Geoffrey; Williams, Ronald (9 October 1986). "Learning representations by back-propagating errors"(PDF). *Nature*. 323: 533–536. Bibcode:1986Natur.323..533R. doi:10.1038/323533a0.

[12] Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression" (PDF). *The American Statistician*. 46 (3): 175–185.

[13] Matiisen, Tambet (December 19, 2015). "Demystifying Deep Reinforcement Learning". *neuro.cs.ut.ee*. Computational Neuroscience Lab. Retrieved 2018-04-06.

[14] Cortes, Corinna; Vapnik, Vladimir (September 1995). "Support-vector networks". *Machine Learning*. Kluwer Academic Publishers.

[15] Victor Roman. "Aprendizaje Supervisado: Introducción a la Clasificación y Principales Algoritmos". Medium, 27 Març 2019. [Online] Available: https://medium.com/datos-y-ciencia/aprendizaje-supervisado-introducci%C3%B3n-a-la-clasificaci%C3%B3n-y-principales-algoritmos-dadee99c9407. [Accessed: 12 June 2019]

[16] *"The Netflix Prize Rules". Netflix Prize. Netflix. Archived from the original on 3 March 2012. Retrieved 16 June 2016.*

[17] Markoff, John (17 February 2011). "Computer Wins on 'Jeopardy!': Trivial, It's Not". *New York Times*. p. A1. Retrieved 5 June 2016.

[18] "Google achieves AI 'breakthrough' by beating Go champion". *BBC News*. BBC. 27 January 2016. Retrieved 5 June 2016.

[19] Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanisław Jastrz˛ebski ,Thibault Févry, Joe Katsnelson, Eric Kim, Stacey Wolfson, Ujas Parikh, Sushma Gaddam, Leng Leng Young Lin, Kara Ho, Joshua D. Weinstein, Beatriu Reig, Yiming Gao, Hildegard Toth, Kristine Pysarenko, Alana Lewin, Jiyon Lee, Krystal Airola, Eralda Mema, Stephanie Chung, Esther Hwang, Naziya Samreen, S. Gene Kim, Laura Heacock, Linda Moy, Kyunghyun Cho, Krzysztof J. Geras. “Deep Neural Networks Improve Radiologists’ Performance in Breast Cancer Screening”. [Online] Available: https://aiforsocialgood.github.io/icml2019/accepted/track1/pdfs/10_aisg_icml2019.pdf. [Accessed: 3 September 2019]

[20] Adarsh Verma. “Feature Extraction from Text (text data preprocessing)”. Medium, 100-Days-of-ML-and-Code, 27 Abril 2019. [Online] Available: https://medium.com/100-days-of-ml-and-code/feature-extraction-from-text-text-data-preprocessing-594b11af19f5. [Accessed: August 18 2019]

[21] Intuition Engineering. “Deep learning for specific information extraction from unstructured texts”. Medium, 21 July 2018. [Online] Available: https://towardsdatascience.com/deep-learning-for-specific-information-extraction-from-unstructured-texts-12c5b9dceada. [Accessed: 19 April 2019]

[22] Carletta, Jean. (1996) Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics, 22(2), pp. 249–254.

[23] Sarang Narkhede. “Understanding AUC – ROC curve”. Medium, 26 Juny 2018. [Online] Available: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5. [Accessed: 25 August 2019]

[24] Powers, David M W (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (PDF). *Journal of Machine Learning Technologies*. 2(1): 37–63.

[25] Wilame Lima Vallantin. “Why is removing stop words not always a good idea”. Medium, 22 Gener 2019. [Online] Available: https://medium.com/@wilamelima/why-is-removing-stop-words-not-always-a-good-idea-c8d35bd77214. [Accessed: 18 August 2019]

Glossari

ML: Machine Learning

SL: Supervised Learning

RL: Reinforcement LEarning

NLP: Natural Language PRocessing

NCBI: National Center for Biotechnology Information

ICO: Institut Català d'Oncologia

TFIDF: Time Frequency – Inverse Document Frecuency

BoW: Bag of Words

SVM: Supported Vector Mahcines

HTTP: Hyper Text Transport Protocol

CV: Cross-Validation